Statistical inference with the GSS data

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

```
load("gss.Rdata")
```

---

1. Introduction

The General Social Survey (GSS) is an observational study conducted by the National Opinion Research Center, at the University of Chicago.

Since 1972, the GSS has been monitoring sociological and attitudinal trend, by gathering data on contemporary American society. The purpose is to explain trends and constants in attitudes, behaviors, and attributes. From 1972 to 2004, the GSS' target population was adults (18+) living in households in the United States, who were only able to do the survey in English. However, from 2006 till date, it has included those able to do the survey in English or Spanish.

Before 1994, GSS was conducted almost annually (except in 1979, 1981, or 1992, due to funding limitations). Since then, the GSS has been conducted in even numbered years, using dual sample design. This is done, majorly, via face-to-face interviews. In 2002, GSS started Computer-assisted personal interviewing (CAPI). Also, at times when it has proved difficult to arrange an in-person interview with a sampled respondent, interviews have been conducted via telephone.

From 1972 to 1974 surveys, modified probability sampling (block-quota sample) was used. Full-probability sampling of households was used by GSS, to give each household an equal probability of being included in the survey, from 1975 to 2002. As a result of this, the GSS is self-weighting for household-level variables . In order to keep the design unbiased, GSS started to use a two-stage sub-sampling design for nonresponse and weights adjusted, from 2004. Cases from which no response was received after the initial stage of the field period are subsampled, and resources are focused on gaining cooperation from this subset, thus reducing both response error and nonreponse bias. The subsampling of segments was done using a simple systematic selection procedure.

Weight variables are included to adjust the alterations in sampling e.g over-representation of blacks in 1972, under-representation of male for all full-probability samples, under-representation of men in full-time employment for block-quota samples; and significant increment in the coverage of Mormons when the 1980 sample frame (controlled selection procedure in the first stage) was adopted (this was due to the addition of a primary sampling unit in Utah).

The GSS is a result of retrospective observational study (as the data recorded are events that have taken place) and not an experimental study, hence no random assignment. We cannot make causal conclusions from the data (we can only associate). GSS uses random sampling as explained above; hence the data is generalizable. We are simply saying that since there is no random assignment, only random sampling, there is no causal relationship but only association and the data is generalizable.

Our analysis will focus on the GSS 2012 report; hence, this data is generalizable to adults, aged 18 years and above, living in households in the United States, who are able to do the survey in English or Spanish.

---

2. Research question 1 :

Is there any relation between political party affiliation and race to which people belong?

2.1 Motivation:

As American Presidential elections are going on ,it would be of great importance to know how votebank is influenced by different races. This might change

---

2.2 Exploratory data analysis

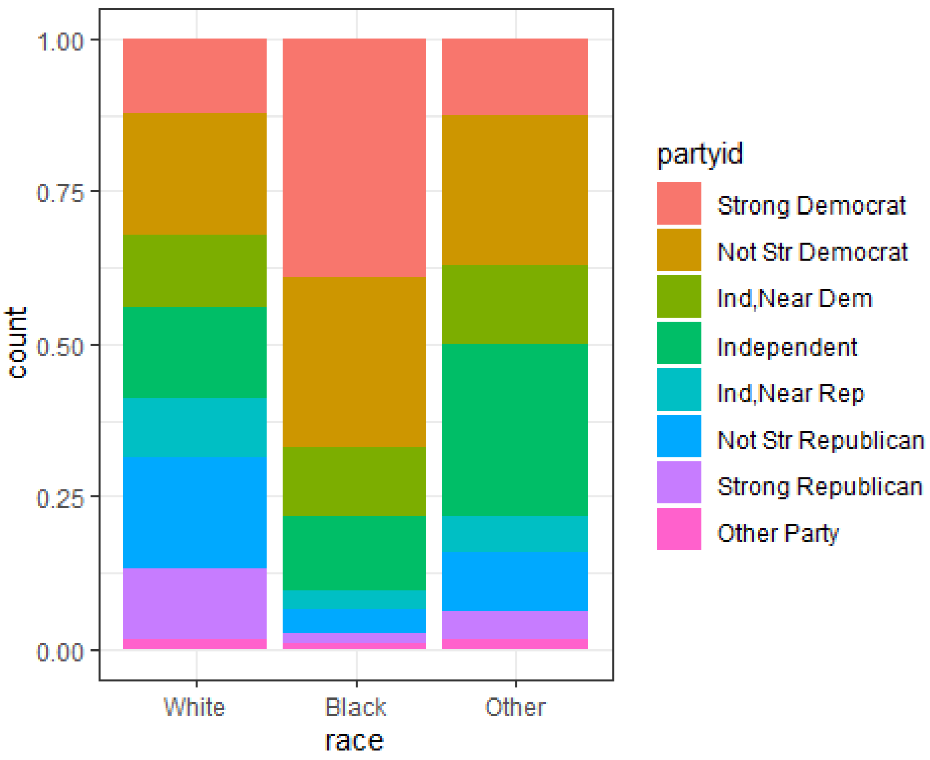
```
#sum(is.na(gss$partyid))
gss%>%select(race,partyid)%>%filter(!is.na(partyid))->new_df
View(new_df)
```

Visualizing different race affiliation to political parties in US.

```
#View(new_df)
sum(is.na(new_df))

## [1] 0

ggplot(new_df, aes(x = race, fill = partyid)) +
 geom_bar(position = "fill")+theme_bw()
```

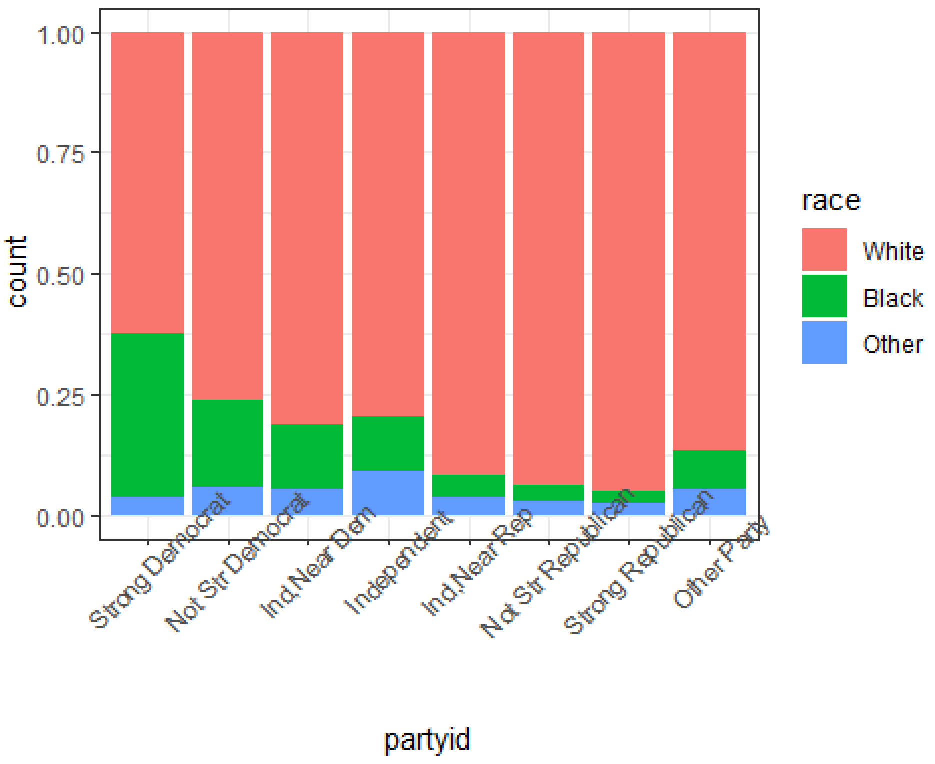


Interpretation:

If we consider different races and their political affiliation, *white* race has somewhat equal proportion of political affiliation ,but at the same time it has maximum and of all races greatest no of *Strong Republican* affiliation. Similarly, Black population seem to have *Strong Democrat* affiliation and least *Strong Republican* affiliation.This is exactly opposite to *White* population. Others have roughly equal political party affiliation.

For more insight.

```
ggplot(new_df, aes(x = partyid, fill = race)) +
 geom_bar(position = "fill")+theme_bw()+theme(axis.text.x  =
element_text(angle = 45))
```



2.3 Inference:

2.3.1 Stating Hypothesis:

Null hypothesis: Race and Partyid are independent of each other.

Alternative hypothesis: Race and Partyid are dependent.

2.3.2 Checking Conditions

Independence: As mentioned above, samples are collected based on random sampling method. The sample size (57,061) is less than 10% of the population (population of the US). Therefore, the condition of independence between groups is guaranteed.

Expected Counts:

```
chisq.test(new_df$partyid,new_df$race)$expected
```

```
##                     new_df$race
## new_df$partyid           White     Black     Other
##   Strong Democrat    7412.6446 1261.7951 442.56033
##   Not Str Democrat   9789.2114 1666.3391 584.44954
##   Ind,Near Dem       5482.4462  933.2329 327.32087
##   Independent        6910.1751 1176.2638 412.56118
##   Ind,Near Rep       4001.0556  681.0676 238.87676
##   Not Str Republican 7321.5821 1246.2943 437.12359
##   Strong Republican  4510.8426  767.8446 269.31279
##   Other Party         700.0424  119.1626  41.79494
```

From the above table, the expected counts are above the minimum required of 5 for each cell.

Degrees of Freedom: The degrees of freedom is given by 14 (= (3-1)*(8-1)).

All the conditions to perform chi-square test of independence are satisfied.

2.3.3 Carrying out Chi-Sq Test:

```
chisq.test(new_df$partyid,new_df$race)

## 
##  Pearson's Chi-squared test
## 
## data:  new_df$partyid and new_df$race
## X-squared = 5670.8, df = 14, p-value < 2.2e-16
```

Here *p-value* is nearly 0 and in smaller than level of significance (alpha= 0.05), threrefore we reject Null Hypothesis.

Conclusion 2.3.4:

We have enough evidence to say that there is association between race of population and party affiliation.

---

3 Research question 2:

Is there income disparity among different races in US society?

3.1 Motivation

The reason why this question is interesting is that this research surveys across a broad range of US demographics, including all races. Studying the differences between average family income across races may reveal helpful insights about the income gap in the US.

3.2 Exploratory data analysis

First, we will start with exploring different facets of our family income data. We start with the measure of center, specifically the mean. We want to find out what is the average family income of people conducting this survey. We exclude NA answers in our computation.

```
mean(gss$coninc, trim = 0, na.rm = TRUE)

## [1] 44503.04
```

The result shows that the average family income is 44,503.04 (rounded to 44,504) US dollars.

Next, we will examine the median value. Again, we exclude NA answers in our computation.

```
median(gss$coninc, trim = 0, na.rm = TRUE)

## [1] 35602
```

The median value is 35,602 US dollars. There is quite a big difference between the mean and the median values.

Coming to the measure of spread, we will calculate the standard deviation

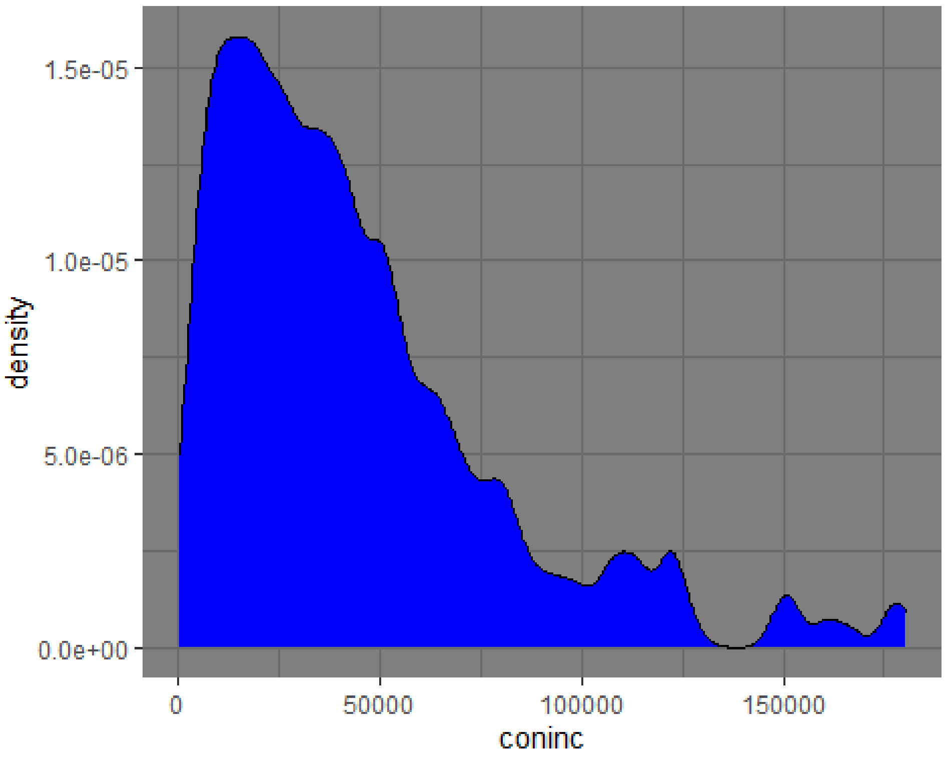
```
sd(gss$coninc, na.rm = TRUE)

## [1] 35936.01
```

The standard deviation is 35,936.01. The data is quite spread around the mean.

We will visualise this spread using density curve.

```
gss%>%select(coninc)%>%ggplot(aes(coninc))+geom_density(fill =
"blue")+theme_dark()
```

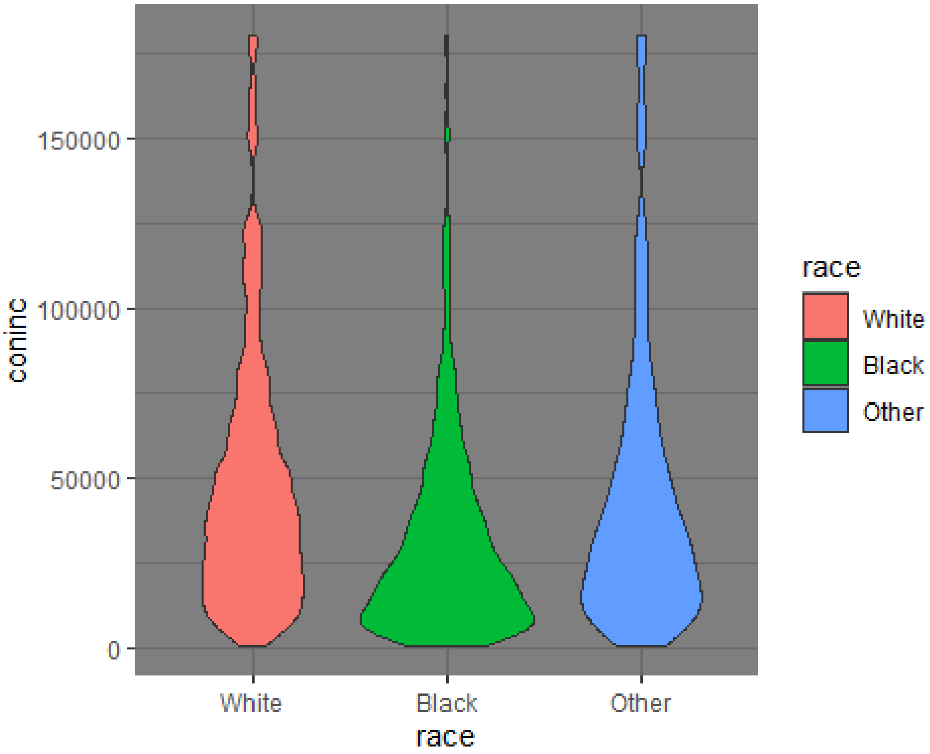


Interpretation:

It can be seen that income is unimodally distributed in US.

Now we will try to visualise this income distribution among different races using violine plot

```
gss%>%select(race,coninc)%>%ggplot(aes(race,coninc,fill =
race))+geom_violin()+theme_dark()
```



Interpretation:

Here we can see that, White & Other races have similar kind of income distributions. And greater proportion of the people belonging to Black race have fairly low income. As interpretation of graphs/plots is subjective, we try to confirm this hypothesis using Statistical tests.

3.3 Inference:

3.3.1 Stating Hypothesis:

μ = average family income (in US dollars)

Null hypothesis: The average family income is the same across all races(i.e μ1 = μ2 = μ3).

Alternative hypothesis: The average family income differs between at least one pair of races.

3.3.2 One-way ANOVA

We will use a one-way ANOVA test in this case. This is because we have 1 independent variable (family income in US dollars). We want to compare the differences in the average family income between 3 races (White, Black, Other). The number of observations is not the same between those 3 groups.

How to perform the test? Using the aov function, we can observe the sum of squares total (SST), sum of squares group (SSG), sum of squares error (SSE), degrees of freedom, mean squares, F statistic, and p-value.

3.3.3 Checking Conditions

Independence

Within groups

As mentioned above, samples are collected based on random sampling method. The sample size (57,061) is less than 10% of the population (population of the US). Therefore, the condition of independence between groups is guaranteed.

Between groups

As this is individual survey, dependence between groups is not likely to happen. For instance, people with different races or education levels do not have any sign to be dependent to each other.

We can conclude that the independence conditions (both within and between groups) are ensured.

Approximate normality

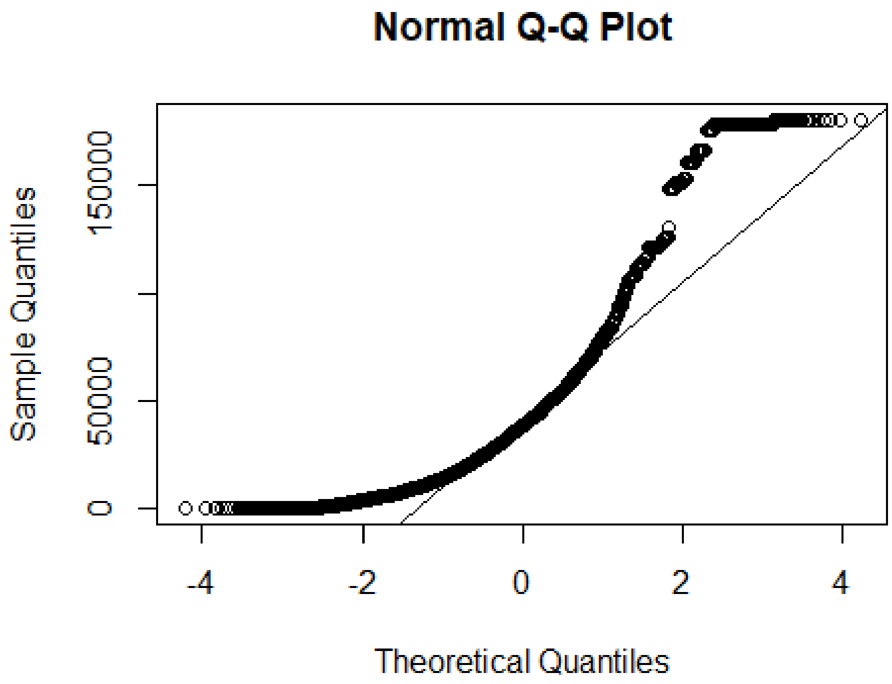
To verify this condition, we draw normal Q-Q plots.

First, we start with the White race by creating a subset:

```r
white <- subset(gss, race == "White")
```

Let's see a Q-Q plot:

```r
qqnorm(white$coninc)
qqline(white$coninc)
```



Interpretation:

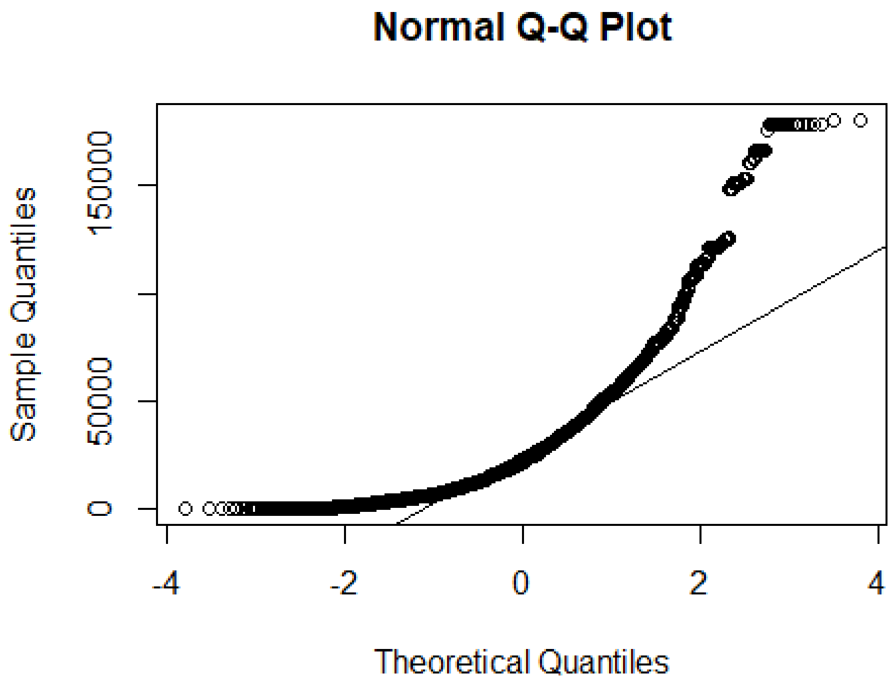
We can see that distribution of response variable (White group) looks approximately normal.

Similarly we will plot Q-Q Plot for other races.

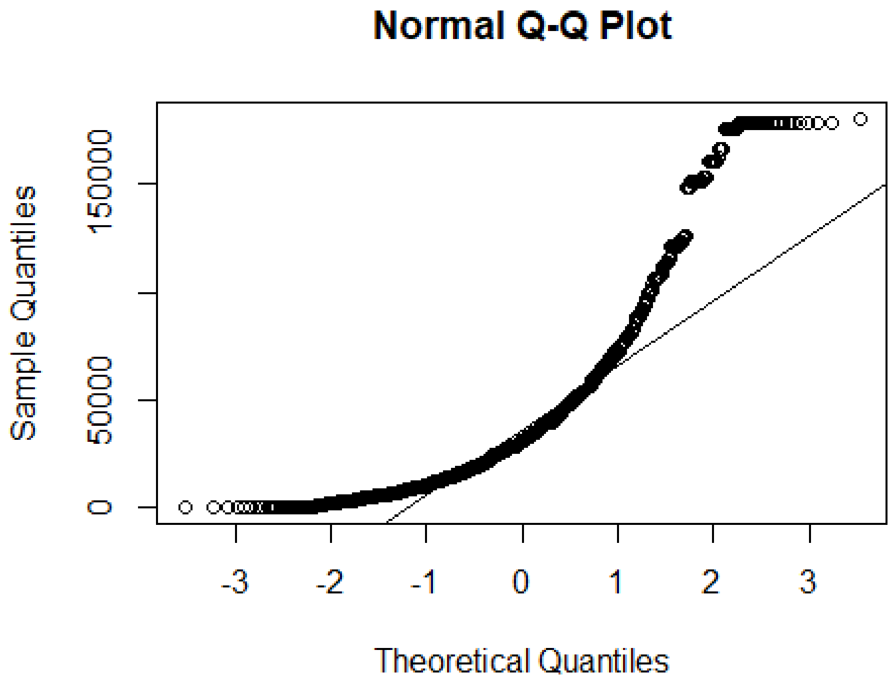
```r
black <- subset(gss, race == "Black")
others <- subset(gss, race == "Other")
```

Q-Q plot:

```r
qqnorm(black$coninc)
qqline(black$coninc)
```



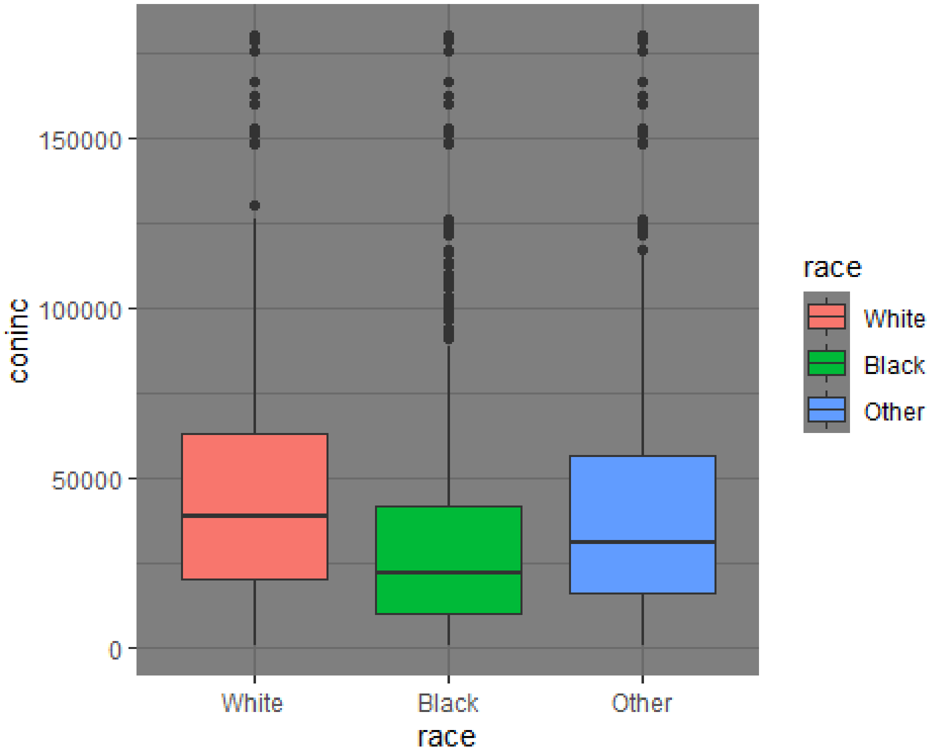
```
qqnorm(others$coninc)
qqline(others$coninc)
```



We can see that distribution of response variable (White,Black & Other group) looks approximately normal.

Constant Variance

```
gss%>%select(race,coninc)%>%ggplot(aes(race,coninc,fill =
race))+geom_boxplot()+theme_dark()
```



Interpretation:

Here we observe that variability is constant across the groups, therefore this condition is also satisfied.

3.3.2 Carring Out ANOVA

Let's fit the model by the ANOVA formula:

```
aov.out = aov(coninc ~ race, data=gss)
summary(aov.out)

##                Df    Sum Sq   Mean Sq F value Pr(>F)    
## race            2 1.699e+12 8.494e+11   675.1 <2e-16 ***
## Residuals   51229 6.446e+13 1.258e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5829 observations deleted due to missingness
```

Getting through Analysis & Results

```
income_race = lm(gss$coninc ~ gss$race -1, data = gss)
summary(income_race)

##
## Call:
## lm(formula = gss$coninc ~ gss$race - 1, data = gss)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46624 -25028  -8593  14714 150201
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## gss$raceWhite  47006.7      173.5  271.01   <2e-16 ***
## gss$raceBlack  30185.0      425.3   70.97   <2e-16 ***
## gss$raceOther  42415.4      716.4   59.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35470 on 51229 degrees of freedom
##   (5829 observations deleted due to missingness)
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.6154
## F-statistic: 2.733e+04 on 3 and 51229 DF,  p-value: < 2.2e-16
```

Interpretation:

As p-value is much smaller than level of significance(l.o.s) i.e(2.2e-16 << 0.05, therefore we reject null hypothesis.

3.3.4 Conclusion:

The data provide convincing evidence that at least one pair of population means are different from each other. There is a difference between the average family income (is US dollars) between at least one pair of race.